overSEAS 2025

This thesis was submitted by its author to the School of English and American Studies, Eötvös Loránd University, in partial fulfilment of the requirements for the degree of Bachelor of Arts. It was found to be among the best theses submitted in 2025, therefore it was decorated with the School's Outstanding Thesis Award. As such it is published in the form it was submitted in **overSEAS 2025** (http://seas.elte.hu/overseas/2025.html)

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

Faculty of Humanities

BA THESIS

Mediation tasks in Hungarian state-accredited bilingual language examinations: An exploratory study

Magyarországi akkreditált kétnyelvű vizsgarendszerek közvetítési feladatainak feltáró vizsgálata

Written by:

Bajáki Bence

English and American Studies

English track

2025

CERTIFICATE OF RESEARCH

By my signature below, I certify that my ELTE B.A. thesis, entitled *Mediation tasks in Hungarian state-accredited bilingual language examinations: An exploratory study*, is entirely the result of my own work, and that no degree has previously been conferred upon me for this work. In my thesis I have cited all the sources (printed, electronic or oral) I have used faithfully and have always indicated their origin. The electronic version of my thesis (in PDF format) is a true representation (identical copy) of the printed version. If this pledge is found to be false, I realize that I will be subject to penalties up to and including the forfeiture of the degree earned by my thesis.

Date: 2025. 04. 15. Signed: Bajáki Bence

EÖTVÖS LORÁND TUDOMÁNYEGYETEM

Bölcsészettudományi Kar

ALAPSZAKOS SZAKDOLGOZAT

Mediation tasks in Hungarian state-accredited bilingual language examinations: An exploratory study

Magyarországi akkreditált kétnyelvű vizsgarendszerek közvetítési feladatainak feltáró vizsgálata

Témavezető:

Dr. Tankó Gyula

Habilitált egyetemi docens

Készítette:

Bajáki Bence

Anglisztika alapszak

Angol szakirány

2025

Table of contents

Abstract

In Hungary, there are two state-accredited bilingual language examinations that involve tasks which require the mediation of L2 texts into L1 and issue L2 certificates partially on the basis of mediation tasks. The question naturally presents itself: How do such tasks provide information necessary to make decisions about L2 proficiency? To answer this question, this study reports the analyses of two such L2àL1 mediation tasks. The results suggest that the measurement of L2 is insufficient with these tasks, and additionally, the analyses revealed some general concerns about the way the tasks are designed, which calls into question the validity of these examinations.

Keywords: *language assessment, mediation, bilingual examination*

1. Introduction

According to the Common European Framework of Reference for Languages (Council of Europe, 2001, 2020), mediation is primarily a language activity by means of which communication is made possible between language users for whom this would otherwise be impossible due to their inability to communicate directly. In language examinations, together with the language abilities measured by other tasks, mediation tasks are expected to provide a more comprehensive account of test takers' overall language proficiency (Benke, 2002). In Hungary, several bilingual language examinations exist, which include tasks whose aim is to measure test takers' ability to mediate between two languages (Educational Authority, 2025a). Such examinations can test general language proficiency (e.g., Origo, Euro, Ezra, KJE) or language proficiency for specific purposes (e.g., ARMA, Profex, BGE).

The mediation-based tasks typically require test takers to produce an oral or written text in the assessed foreign language. There are, however, language examinations that assess foreign language proficiency with tasks that require the translation of an L2 text into the test takers' L1, so these examinations assess L2 proficiency partly based on the evaluation of an L1 text. This poses a problem as it is not clear what information such tasks provide for making L2 certification decisions (Bachman & Palmer, 2010).

This study therefore aims to explore what specific abilities and in what way Hungarian state-accredited bilingual specific purpose language examinations test by means of mediation tasks that involve translating L2 texts into L1. As there is hardly any literature that specifically deals with the issue of assessing L2 based on L1 performance, to answer this question, two translation task types are analysed that entail translating English texts into Hungarian. The tasks are taken from two Hungarian state-accredited bilingual specific purpose language examinations. Following a literature review of the concepts needed for analysis, the

methodology of the research is described. Finally, the analyses of the examination tasks are presented before some conclusions are formulated.

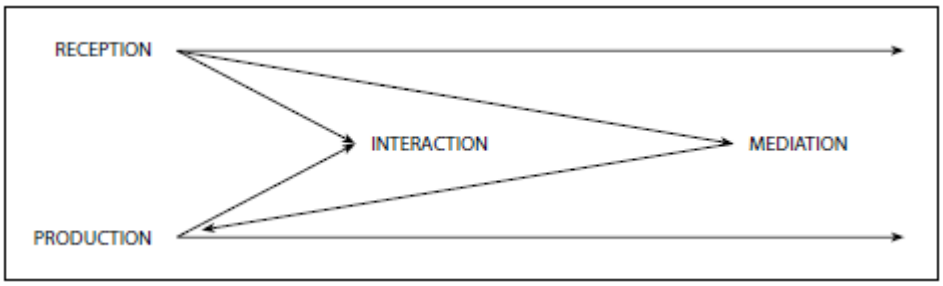
2. Literature review

2.1 Mediation

According to the Common European Framework of Reference for Languages (CEFR, Council of Europe, 2001, 2020), mediation is primarily a language activity by means of which communication is made possible between language users for whom this would otherwise be impossible due to their inability to communicate directly. The CEFR differentiates four communication activities (see Figure 1): reception, production, interaction, and finally, mediation, which can incorporate the other three activities (Council of Europe, 2001, 2020).

Figure 1

The relationship between reception, production and mediation (Council of Europe, 2020, p. 34)



Jakobson (1959) distinguished between three types of mediation: intralingual, interlingual, and intersemiotic. Intralingual mediation involves the activities that mediate meaning within the bounds of the same language, its most common forms being paraphrasing and summarising used separately or in combination (Tankó, 2022). Interlingual mediation covers those activities that involve the transfer of meaning between two different languages, and the most common types are translating and interpreting. Intersemiotic mediation extends

the previous two concepts to the mediation of meaning that includes non-verbal signs (i.e., audiovisual translation).

Another important aspect of the types of mediation to consider are the modes in which they happen. The Common European Framework of Reference for Languages (Council of Europe, 2001) differentiates between written and oral mediation, or translation and interpretation. Kaindl (2013) made a further distinction with the concepts of intramodal and intermodal mediation. Intramodal mediation covers mediation in which the modes of the reception and production are the same, as is typical in translation (i.e., text-to-text) and interpretation (i.e., speech-to-speech). However, in intermodal mediation, these activities happen across modes as, for example, in the case of the oral retelling of a written text. The examination tasks analysed in this paper involve the interlingual written translation of a written text (Profex examination) and the interlingual oral summarisation of a written text (BGE examination).

2.2 Mediation in language assessment

According to Benke (2002), the validity of mediation tasks in foreign language examinations has been a debated issue for some time. Therefore, to uncover how mediation measures the aspects of language ability assessed by conventional tasks and whether this causes redundancy in the assessment, and also whether it measures aspects additional to those, Benke conducted an intersubtest- and whole test-subtest correlation study in which written language examination tests consisting of the conventional reading and writing subtests and a mediation subtest were administered to university students of statistically significant sample sizes on two occasions. In both instances, the correlation values fell into the normal range, indicating that the aspects of language ability measured by the mediation subtest constitute a part of the language ability measured by the other subtests, and measures them in a similar way. However,

it does not cause redundancy, according to Benke (2002); therefore, it contributes to the reliability of the examinations, as it broadens the sample of and extends the variety of the means of assessment.

2.3 Concepts necessary for the analysis

2.3.1 Bilingual and technical examinations

In Hungary, a state-accredited bilingual complex language examination is an examination that measures all of, though not exclusively, the following five language abilities: listening comprehension; oral competence; writing; reading comprehension; mediation (Educational Authority, 2025b).

2.3.2 Construct

Bachman and Palmer (1996) define the term "construct" in language assessment, as the ability or set of abilities that is intended to be assessed. The construct definition is the foundation based on the assessment is developed, and also provides brief yet comprehensive information about the assessment to stakeholders.

2.3.3 Criteria for correctness

An assessment's criteria for correctness specify what constitute correct responses to a task; they are the aspects of language ability based on which the assessment's construct is measured (Bachman & Palmer, 2010). For the analytic purposes of this paper, Bachman and Palmer's binary concept of the criteria (correct/incorrect) is expanded to classify the criteria as scored on multiple levels of the ability specified by each criterion.

2.3.4 Holistic and analytic scales

Bachman and Palmer (2010) defined the concepts of holistic and analytic rating scales as follows: holistic (or global) scales treat language ability as single and unitary, define degrees

of said ability as a mix of varying definitions of levels of more specific abilities, and assign these to levels of a top-down scale with each level yielding a single numeric score. Analytic scales, on the other hand, take the construct intended to be measured by the assessment, and build individual scales upon each of the different aspects of language ability defined by the construct, with each scale returning a separate score. Bachman and Palmer advised the use of analytic scales, as holistic scales suffer from a subjectivity of interpretation.

2.3.5 Extended and limited production response test item types

Extended production response test item types (Bachman & Palmer, 2010) are items that do not have a set of possible answers to choose from and require the test taker to formulate their own response in the form of a lengthy text (i.e., a letter, a composition, or a translation). Therefore, the range of acceptable responses is wide. Limited production response items require candidates to produce a word, phrase, or the in the case of the current study the translation of a few consecutive words taken from the source text.

2.3.6 Partial credit scoring

Assessments that use partial credit scoring (Bachman & Palmer, 2010) as their scoring method measure multiple criteria for correctness per response and may return partial scores based on the number of criteria satisfied by an answer.

2.3.7 Compensatory and non-compensatory scoring

Bachman and Palmer (2010) differentiated between composite scoring methods based on whether they are compensatory or non-compensatory. In compensatory scoring, scores from multiple separate outputs (e.g., different papers of an examination or sub-scales within an analytic scale) are simply added together to return a single composite score for the given task, so that a “high” score on one scale/paper is added to a “low” score on another, which results in an “average” composite score. In non-compensatory scoring, however, there is a minimum

score assigned to each of the separate outputs, which, if not met even once, will render the composite score null, regardless of the scores returned by the other outputs. These two methods may be applied in the scope of a single task, a subtest of an examination, or the entirety of an examination.

2.3.8 Stochastic independence

As defined by Fulcher and Davidson (2007), the principle of stochastic independence states that the extent of correlation between outcomes on different scoring instruments exponentially decrease the amount of unique information provided by each outcome, deteriorating the quality of the assessment; therefore, correlations between outcomes may not be desirable in testing, especially if the correlation is due to lack of independence between test items.

2.3.9 Assessment Use Argument

According to Tankó (2019)'s account, Bachman and Palmer (2010)'s Assessment Use Argument model is based on Toulmin (1958, 2003)'s argumentation model, and is one of the most widely used frameworks for scientifically measuring the validity of assessments. The model is a system of arguments that articulates four claims that build onto each other: (i) the assessment records of student performances (scores, grades, etc.) are the basis for the (ii) interpretations about their language abilities, on which (iii) decisions are made that have (iv) consequences. Each claim has a number of warrants that articulate principles in support of the claims. A claim is satisfied when all the principles in its warrants are backed, and an assessment is only considered valid if all four claims are substantiated.

Using the concepts discussed in the review of the literature, this study aims to explore what specific abilities and in what way Hungarian state-accredited bilingual specific purpose

language examinations test by means of mediation tasks that involve translating L2 texts into L1.

3. Research Methods

In this study, an analysis was conducted on two translation tasks that require the translation of L2 (English) texts into L1 (Hungarian). The tasks originate from two Hungarian state-accredited bilingual English for specific purposes examinations. The two examinations are (i) the B1-level English for Medical Purposes examination administered by the Faculty of General Medicine, University of Pécs (PROFEX) and (ii) the C1-level tourism and catering examination administered by the Foreign Language Examination Centre of Budapest Business University (BGE). Both examinations include a task that requires the translation of a written English text into Hungarian. The analysis was conducted based on the publicly available sample tasks and the documentation provided for these two tasks, which are listed below:

(i) PROFEX examination (see Appendix A)

- Rating guide
- Sample task

(ii) BGE examination (see Appendix B)

- Rating guide
- Sample task

The analysis was based primarily on concepts defined by Bachman and Palmer (2010) in the AUA framework and on my own inferences (i) due to the lack of literature on mediating to L1 in foreign language assessment, and also (ii) because these mediation tasks do not fit perfectly into any existing framework. Although the abundance of differences between the two tasks and examinations makes the tasks unfit for parallel analysis, the uniqueness of these tasks in the broader language assessment context and the lack of former research on the subject warrant the analysis of both.

Both analyses start with the general descriptions of the tasks, followed by the formulation of the construct definitions based on the rating guides, sample tasks, sample solution, and any other available documentation. Next, the scoring instruments contained in the two rating guides—which are in Hungarian and show substantial differences in terms of both structure and content—had to be translated to English. In order to make them more comprehensive and comparable, the scoring instruments had to be restructured into scales that are similar to each other in terms of organisation and logic. Some pieces of information needed for the restructured scoring instruments were not explicitly present in the rating guides; therefore, these needed to be supplemented by way of deduction from context. During the translation of the scoring instruments, the emphasis was placed on translating the descriptors to resemble as closely the wording of the original as possible; however, in the restructured scales, in some instances, the wording was tweaked in the interest of meaningfulness and analysability. Simultaneously with, and after the translation and restructuring processes, general analyses of the scoring instruments are conducted based on concepts listed in the Literature Review section or articulated in the analyses. Finally, the criteria for correctness derived from the scoring instruments are individually analysed for both tasks in terms of whether and how effectively they measure L2 proficiency from an L1 output.

4. Analysis of the Profex examination mediation task

The mediation subtest, as part of the written examination in the Profex B1-level English for Medical Purposes examination, consists of a task that requires the test taker to translate a 200–250 word long written English text into a written Hungarian text of similar length. The rating is carried out with two separate scoring instruments: a holistic scale measuring the "overall impression" of the produced translation on which zero to five points can be awarded, and a table for scoring the translation of five specific segments of the source text, on which

zero to two points are awarded per segment. These add up to a maximum composite score of fifteen points of which a minimum of 6 points (40%) need to be achieved in order to pass the mediation subtest. No explicitly formulated official construct definition is publicly available for the task. For this reason, one had to be formulated based on the available documentation, which states that the construct the task measures is the "ability to translate simple, coherent L2 professional texts to L1" (cf. Warta et al., 2025). as it is not evident how interpretations of the test taker's L2 (English) proficiency can be made based on a text produced in their L1 (Hungarian). To uncover in what way, and how efficiently the task in question assesses the construct, and crucially, to what degree this construct is relevant (Bachman & Palmer, 2010) and justifiable for making L2 certification decisions, first, the criteria for correctness had to be established. Since the descriptions provided by the rating guide are not sufficiently structured and comprehensive, an attempt was made to restructure and supplement them in the interest of better analysability. Since the two scoring instruments are fundamentally different, their restructuring and analysis are conducted separately prior to the holistic analysis of the task as a whole. Issues and concerns that are not directly related to L1-based L2 assessment are addressed in the analysis as these are essential to provide a comprehensive analysis of the task.

4.1 Restructuring and analysis of the scoring instruments

The aim of the restructuring is to transform the two scoring instruments, the "overall impression" holistic scale, and the descriptions of the rating of the "highlighted parts" into two scales that are comparable in terms of structure and content. The primary feature of the restructured scales is that they separate the original descriptions to establish the criteria for correctness and provide a comprehensive set of the levels of ability assigned to them.

4.1.1 The holistic scale

As the original rating guide is written in Hungarian, a translated version of the "overall impression" holistic scale in its original state is provided in Figure 2. With this scale, zero to five points can be awarded on six levels As the original rating guide is written in Hungarian, a translated version of the "overall impression" holistic scale in its original state is provided in Figure 2. With this scale, zero to five points can be awarded on six levels (0–5) of language ability, based on the overall impression of the test taker's performance. The scale is inconsistently and poorly structured, the descriptions are formulated in an arbitrary number of half-sentences per level, each of which contains an arbitrary number of criteria. To derive the criteria for correctness and make the holistic scale resemble better structured and more transparent analytic scales, the scale was divided into separate columns matching the number of the criteria present in the original.

Figure 2

Translation of the Profex examination holistic scale

Overall impression	
5 points	- Use of the correct grammatical structures and adequate vocabulary - Mediation of all important information, if with a few minor deficiencies or inaccuracies
4 points	- Substantive meaning of the text mediated with minor inaccuracies, a few grammatically incorrect structures or lexical inaccuracies - Minor deficiencies in information that do not violate cohesion
3 points	- Frequent inaccuracies in the translation - A few major and frequent minor grammatical mistakes - Cohesion violated on 1-2 instances
2 points	- Frequent violation of cohesion - Frequent mistranslations and inaccuracies
1 point	- The text barely contains any assessable, intelligible or correct information
0 points	- Performance not assessable

For some levels (marked with an asterisk in Figure 3) of the criteria, the descriptions were not explicitly present in the original; therefore, they needed to be assumed based on their diagonal adjacents.

Figure 3

Criteria for correctness derived from the Profex examination holistic scale

<table>
<tr><th colspan="5">Overall impression</th></tr>
<tr><th rowspan="2">Score</th><th colspan="4">Criteria for correctness</th></tr>
<tr><th>Grammatical correctness</th><th>Cohesion</th><th>Completeness</th><th>Lexical precision</th></tr>
<tr><td>5</td><td>all grammatical structures are correct</td><td>cohesion not violated*</td><td>all important information mediated, minor inaccuracies or deficiencies allowed</td><td>adequate vocabulary</td></tr>
<tr><td>4</td><td>a few grammatical structures are incorrect</td><td>cohesion not violated</td><td>gist of important information mediated, minor inaccuracies or deficiencies allowed</td><td>few lexical imprecisions</td></tr>
<tr><td>3</td><td>a few major and frequent minor grammatical mistakes</td><td>cohesion violated at 1-2 instances</td><td colspan="2">frequent inaccuracies</td></tr>
<tr><td>2</td><td>major mistakes frequent*</td><td>frequent violation of cohesion</td><td colspan="2">frequent inaccuracies and mistranslations*</td></tr>
<tr><td>1</td><td colspan="2">text barely intelligible</td><td colspan="2">barely any relevant information present</td></tr>
<tr><td>0</td><td colspan="4">performance not assessable</td></tr>
</table>

Since the descriptions in the horizontal rows of the scale are assessed together and are assigned one score, each description must naturally assign itself a logical relationship (and/or) in relation to those with which it shares a row so as to be able to give a full description of the levels of language ability assigned to its score. However, these horizontal relationships are impossible to deduce from the original with certainty. Additionally, a test taker's performance may represent inconsistent levels on these separate criteria, further increasing the ambiguity of scoring. Although raters often have access to more detailed rater guidelines (Tankó, 2005), as is often the case with holistic scales, establishing these logical relationships or choosing one score to represent differing levels of language ability is often left for the rater's personal judgement, highlighting some major shortcomings that contribute to the problems of intra- and interrater inconsistency, interpretation, and meaningfulness that characterise holistic scales (Bachman & Palmer, 2010).

4.1.2 Specific, separately scored stretches of the translated text

Ten additional points could be awarded for the quality of the translations of five highlighted parts in the translation, that is for specific, separately scored stretches of the

source text. This scoring instrument targets what Bachman and Palmer (2010)'s defined as limited production responses with partial credit scoring and consists of the descriptions of the conditions for point deduction. A rating sample that demonstrates examples of possible explanations for point deduction. As was the case with the "overall impression" scale, these also had to be translated from Hungarian (see Figure 4).

Figure 4

Translation of the Profex examination's "highlighted parts" scoring instrument

Highlighted parts

- 2 points get deducted if it makes no sense in Hungarian, if the information is misunderstood or deficiently mediated, or if text cohesion or coherence is violated
- 1 point gets deducted if sentence composition is clunky, but the mistakes do not affect the essence of the information mediated (e.g., misunderstanding a word or short structure, improper word choice, missing non-essential word)

Rating sample		
Highlighted part	**Score**	Explanation
1.	2	
2.	2	
3.	**0**	The test taker completely misunderstood the information; therefore, its mediation has not been realised.
4.	**1**	The mistake did not affect the essence of the information, but ungrammatical formulation hinders the mediation of information.
5.	2	

The descriptions suffer from the same problems as, and are even less structured than, the ones in the "overall impression" scale; therefore, the rationale for, and the methods of reconstruction are the same, except that the rating guide did not provide a description for a maximum-point execution of the task. Instead, the possible mistakes were listed with the respective number of points to be deducted if committed. As was done for the restructured "overall impressions" scale in Figure 3, the descriptions missing from the original (marked with an asterisk in Figure 5) had to be assumed; however, in this case, they were completely straightforward.

Figure 5

Criteria for correctness derived from the Profex examination's "highlighted parts" scoring instrument

Highlighted parts			
Score	**Criteria for correctness**		
	Grammatical correctness	**Lexical precision**	**Cohesion & coherence**
2	*perfect syntactic structure**	*perfect lexical precision**	*cohesion and coherence not damaged**
1	syntactic structure imperfect	minor lexical imprecisions	*cohesion and coherence not damaged**
0	makes no sense in Hungarian	Information misunderstood/insufficiently mediated	cohesion or coherence damaged

Unlike with the "Overall impression" scale, the horizontal logical relationships were deductible from the rating guide: the ones put in italics in Figure 5 assign themselves "and", while the rest assign "or". For example, the level that awards 1 point can be described as "either with minor lexical imprecisions, or has imperfect syntactic structure; and cohesion and coherence are not damaged".

4.1.3 Analysis of further aspects of the Profex examination

Since the highlighted parts are from the same text whose translation the "Overall impression" scale measures and the criteria for correctness (with the exception of completeness) are shared between the two scoring instruments, a mistake in the translation of a highlighted part potentially results in the deduction of points in two instances, thereby violating the principles of stochastic independence and making the assessment unfair. The highlighted parts are not necessarily complete sentences, nor are so their expected translations, which presents another possible issue because under the criterion "grammatical correctness", the majority of the descriptions relate to syntactic correctness, and therefore cannot be observed without considering the highlighted part as part of a complete sentence. A similar problem applies to the criterion of cohesion and coherence, as these concepts apply on a textual level. The task yields a compensatory composite score of fifteen, of which the translation of the

highlighted parts constitutes ten points, which is disproportional to the five points awarded for the translation of the whole text. This is exacerbated by the fact that test takers are not made aware of which specific parts of the source text are highlighted for raters, nor is their existence communicated in the task specification. Given that the translation of the highlighted parts is worth two thirds of the maximum score for the task, it raises issues with the task's adherence to Warrant A2 about the meaningfulness of interpretations in Bachman and Palmer (2010)'s Assessment Use Argument model.

4.2 The criteria for correctness measured

The following criteria for correctness were derived from the two restructured scoring instruments: grammatical correctness; lexical precision; cohesion and coherence; and completeness. These are the aspects of language ability which the test measures, that is the construct described as the "ability to translate simple, coherent L2 professional texts to L1" (cf. Warta et al., 2025). However, as this examination awards certification in English (L2) and not in Hungarian (L1), for it to be relevant and meaningful, its tasks are naturally required to assess proficiency in English above all else. In uncovering how efficient this task is in fulfilling that obligation, an analysis of how each criterion for correctness may measure English proficiency is needed. As these criteria were articulated in the (Hungarian) production domain of language activities, completing the mediation task naturally involves the activity of reception. Therefore, the analysis of the criteria must also pertain to the inferences drawn about reading comprehension.

4.2.1 Lexical precision

Test takers are allowed to use a printed mono- or bilingual general or technical dictionary, making lexical precision a potentially void criterion for measuring reading comprehension or text production either in English or in Hungarian.

4.2.2 Completeness

Mediating all the relevant pieces of information is mainly an indicator of reading comprehension as it depends on the extent to which the test taker understands the source text; however, it cannot be treated as an indicator of ability in terms of production since the language of production is Hungarian, which is not relevant for the L2 certification. Additionally, as in the case of lexical precision, dictionary use also interferes with measuring reading comprehension.

4.2.3 Grammatical correctness

Although there are similarities between Hungarian and English grammar, producing a grammatically correct Hungarian text does not provide substantial evidence of the ability to do the same in English. It can reflect whether the candidate understands a given grammatical structure in English and can render it in Hungarian in a way that the same meaning is expressed (e.g., a conditional English structure expressing impossible condition is translated with the equivalent Hungarian conditional structure; Gy. Tankó, personal communication, March 24, 2025).

4.2.4 Cohesion and coherence

As the coherence of the source text is already established, a simple sentence-for-sentence translation (as is shown in the sample solution provided in Appendix A) would fulfil this criterion without the test taker needing to demonstrate their ability to compose a coherent text. Although the task may be completed with abandoning the coherence set by the source text, and test takers can choose to organise a coherent target text based on their own logic, the fact that this can be circumvented and that doing so is the most likely the conventional solution make the task at best unsuitable for measuring coherence. Although the task may provide proof of the ability to comprehend a coherent English text, and since there are elements in knowledge

of coherence that are not language-specific, proof of the ability to produce coherent texts in Hungarian nonetheless provides no to negligible evidence of the ability to produce coherent texts in English. Although some of the cohesive devices in the Hungarian language function similarly to those in English, and their successful implementation may translate to some degree to the ability to replicate them in English, this does not provide a reliable measurement of that ability. Successful completion of the task may imply that the test taker is familiar with English cohesive devices as part of their reading comprehension skillset; however, allowing dictionary use also interferes here.

4.3 Concluding the analysis of the Profex examination

The results indicate that the mediation subtest suffers from a variety of general issues, deviates from the principles of language assessment, and measures English reading comprehension unreliably, and English production abilities indirectly and to a negligible degree. To provide perspective, the maximum points achievable on the subtest constitute 30% of the maximum points of, and make up half of the points needed for a passing grade on the writing examination, which in itself provides certification, making the task's shortcomings proportionally more severe and having a detrimental effect to the reliability of the examination.

5. Analysis of the BGE examination mediation task

The mediation task in the BGE C1-level tourism and catering examination involves the test taker providing a spoken rendition in Hungarian of a text of 800–1000 keystrokes written in English. This one task constitutes the mediation subtest, and it is part of the oral examination. A maximum composite score of ten points can be awarded for the task on the basis of two separate analytic scales for "mediation" and "reading comprehension", yielding five points each. The ten raw points eventually contribute 20 converted points to the total of 180 points achievable on the oral examination. The website of the examination states that a minimum of

4 raw points (i.e., 8 converted points; 40%) are needed to pass this subtest (BGE Language Examination Centre, 2025); however, in practice, the task uses non-compensatory scoring. As can be seen in Figures 6 and 7, a score of zero to two on any of the two scales results in failing the subtest. Therefore, by all possible permutations, a minimum of 6 raw points (12 converted points; 60%) are needed to pass the mediation subtest. As no official construct definition was publicly available for the task, one had to be formulated based on the rating guide and sample task (see Appendix B). The formulated construct definition states that the construct that this task measures is the "ability to mediate in speech brief, but complex written L2 professional texts to L1". The nature of this construct raises the same questions that the construct of the task in the case of the Profex examination did: How does this construct, and by extension, the task measure L2 (English) proficiency, and whether an examination that contains such a subtest can make reliable certification decisions for issuing a certificate about L2 proficiency. To answer these questions, the criteria for correctness for the task had to be derived from the rating guide. Although on the surface the rating guide consists of two analytic scales, a closer look reveals that they only resemble analytic scales because they are individually scored; however, each contains multiple criteria for correctness. For this reason, in this study they are considered and referred to as holistic scales. Their deconstruction and restructuring reveals the assessed criteria.

5.1 Restructuring and analysis of the scales

As was the case with the Profex task, an English transcription of the original scales was needed for comparison (see Figure 6). In the interest of a comprehensive analysis of the task, issues that do not directly relate to the L1-based assessment of L2 are also included in the analysis.

Figure 6

Translation of the BGE examination's "reading comprehension" and "mediation" scales

	Reading comprehension	**Mediation**
PASS	5 points	
	The test taker comprehends the information in precisely, in detail, and in all its nuances	The test taker produces a linguistically refined, cohesive and coherent Hungarian text with precise technical terms, occasionally with compensatory strategies
	4 points	
	Comprehends the entirety / gist of the information**, mistakes in nuances are present	The test taker produces a cohesive and coherent Hungarian text with mostly accurate technical terms, *with a few omissions of information*
	3 points	
	Correctly comprehends the gist of the information	The test taker produces a linguistically adequate, cohesive and coherent Hungarian text, occasionally uses inadequate or general, instead of technical terms
FAIL	2 points	
	1-2 bigger misunderstandings, omissions are present	The test taker produces an incohesive and incoherent, fragmentary Hungarian text / composition slow, stuttering
	1 point	
	Several severe misunderstandings present, omissions are frequent	*Barely anything was mediated about the text*
	0 points	
	Performance not assessable	Performance not assessable

The scoring instrument consists of two scales, "reading comprehension" and "mediation". In principle, measuring the comprehension and mediation of the same text on separately scored scales violates stochastic independence, as mediation as a language activity incorporates in itself the language activity of reception, or in this case, reading comprehension. However, in this case, the "mediation" scale contains criteria for correctness that unequivocally relate to and describe the language activity of production and do not overlap with the ones that constitute the "reading comprehension" scale, with two exceptions, put in italics in Figure 6. These exceptions are vaguely and ambiguously worded in the original Hungarian, and both relate to information missing from the target text. A straightforward interpretation would assume that they relate to the "completeness" criterion, which is part of the "reading comprehension" scale. Nonetheless, for the assumption that the designers of the task did not create scales that would violate stochastic independence, and for the sake of being able to

reconstruct the scales into consistent and transparent ones, the assumption will be made that the phrases in question mean that the deficiency of information in the target text is due to a lack of the test taker's ability to express said information. This reasoning, however, reveals another concern, namely, that the performance measured by the criteria contained in the "reading comprehension" scale can only be inferred to from the performance measured by the "mediation" scale, and while the majority of possible performances stochastic independence can be observed, a performance classified "level 1" on the "mediation" scale can only draw an outcome of the same level on the "reading comprehension" scale. The same is true for "level 0". However, the severity of this issue is considerably reduced by the fact that due to the non-compensatory scoring method employed by the task, such performances would result in failing the subtest either way; therefore, for such performances, the scores only serve to provide assessment information to stakeholders, and do not unfairly influence certification decisions.

For the restructured scales, separate columns were included for each criterion of correctness, the descriptions not explicitly included for some brackets (marked with an asterisk in Figure 7) were derived on the same logic as in the case of the Profex scales. For better comparability, the two scales are presented together in Figure 7.

Figure 7

Criteria for correctness derived from the BGE examination's "reading comprehension" and "mediation scales"

Mediation		Criteria for correctness				Reading comprehension
Score		Cohesion & coherence	Lexical precision	Linguistic complexity of discourse	completeness	quality
Pass C1	5	cohesive & coherent	lexically precise, a few instances of using compensatory strategies allowed	Linguistically complex	All information present	Precise in all details and nuances
	4	cohesive & coherent	mainly lexically precise	Linguistically adequate*	All information present	Correct, with a few mistakes considering nuances
	3	cohesive & coherent	a few imprecise or overly general expressions	Linguistically acceptable	Gist of information present	Correct comprehension of gist
Fail	2	fragmentary, incohesive, incoherent	a few imprecise or overly general expressions*	Composition slow, stuttering	1-2 major omissions	1-2 major misunderstandings
	1	fragmentary, incohesive, incoherent*	barely anything is revealed about the topic	Composition slow, stuttering*	Frequent omissions	Frequent severe misunderstandings
	0	performance not assessable			performance not assessable	

The horizontal logical relationships of the criteria within the individual scales, as described in the analysis of the Profex task, are similarly inconclusive in this instance, as they were in the case of the Profex holistic scale. Additionally, the possibility of inconsistent ability levels on different criteria shown by the Profex holistic scale is also present in the "mediation" scale, leading to the same conclusions about the unreliability of holistic scales.

The descriptions of the levels of cohesion and coherence treat the criterion as binary and offer no nuances on it. Furthermore, the division between "cohesive and coherent" and "incohesive, incoherent" occurs at the same line that divides "pass" and "fail". This either implies that according to non-compensatory scoring, one mistake in cohesion or coherence results in failing the task, or that more nuanced descriptors are necessary to be able make fair and meaningful interpretations based on this criterion. This is so especially because the mode of production for this task is speech (i.e., discourse less planned than writing), in which mistakes of cohesion or coherence ought to be considered more leniently than in writing.

5.2 The criteria for correctness measured

To separate the varying aspects of language ability based on which the task measures the construct ("ability to mediate in speech brief but complex written L2 professional texts to L1"), the following criteria for correctness were derived during the restructuring of the scales: reading comprehension completeness; reading comprehension quality; linguistic complexity of discourse produced; lexical precision in production; cohesion and coherence in production. To uncover whether, and to what extent, these criteria and by extension the construct can be used to make inferences about L2 (English) proficiency, a one-by-one analysis of the criteria was conducted. As, unlike in the case of the Profex task, these criteria explicitly refer to the language activity domains of reception (reading comprehension) and production, although there is a natural correlation between the two domains, the ways in which the production-based criteria may draw inferences to reading comprehension will not be discussed.

5.2.1 Completeness in reading comprehension

Completeness in reading comprehension assesses how comprehensively the test taker extracts relevant information from the English source text. As dictionary use is not allowed for this subtest, measuring this criterion provides information for relevant and meaningful interpretations concerning the test taker's English proficiency.

5.2.2 Quality of reading comprehension

The criterion quality of reading comprehension measures how precisely the test taker can interpret the English source text, which is phrased in a complex and nuanced manner, in terms of meaning. This criterion provides relevant and meaningful information for interpretations about English language proficiency.

5.2.3 Linguistic complexity of discourse produced

The criterion that measures the extent of linguistic complexity of the Hungarian discourse produced by test takers does not translate to measuring the equivalent of the ability in English, therefore it is not a valid criterion for making relevant and meaningful interpretations on their proficiency in English.

5.2.4 Lexical precision in production

The criterion that measures the extent of lexical precision in the Hungarian discourse produced by test takers does not translate to measuring the equivalent of the ability in English, therefore it is not a valid criterion for making relevant and meaningful interpretations on their proficiency in English.

5.2.5 Cohesion and coherence in production

The criterion that assesses the cohesion and coherence of the Hungarian discourse produced by test takers does not translate to measuring the equivalent of the ability in English, therefore it is not a valid criterion for making relevant and meaningful interpretations on their proficiency in English.

5.3 Concluding the analysis of the BGE examination

As was the case with the Profex task, this subtest also suffers from several issues and does not adequately observe the principles of language assessment. However, despite providing no data resulting in meaningful inferences about production-based language abilities, it measures reading comprehension reliably. This fact, and that this subtest weighs 1:9 in terms of the maximum score of the oral examination, the shortcomings demonstrated by this task are much less severe compared to those of the Profex task.

6. Conclusion

This study aimed to explore how and to what extent mediation tasks that involve translating an L2 text into L1 measure L2 proficiency, and how valid such tasks are for making L2 certification decisions. To that end, separate analyses were conducted of two state-accredited bilingual language examinations from Hungary. The two tasks used for analysis were the Profex B1-level English for Medical Purposes examination, and the BGE C1-level tourism and catering examination. The task in the Profex examination involved translating a written English text into Hungarian, while the task in the BGE examination involved giving an oral account in Hungarian of a text written in English. The analyses uncovered several issues and concerns that were not directly concerned with L2 to L1 mediation, but for the sake of arriving at a comprehensive conclusion, these issues were also included in the discussion.

The analysis of the Profex task showed that the scoring instruments in the rating guide are poorly structured and give overly general and incomplete descriptions of the levels of language ability they intend to measure, and the task specification does not include information crucial for the successful completion of the task, which raises the concern that the examination fails in providing sufficient information to stakeholders. The task also fails to include sufficient information for raters to provide consistent assessments of the test takers' performances. The task also fails considerably in adhering to the principle of stochastic independence, resulting in the possibility of unfairly deducting points on two instances for the same mistake, albeit the same underlying trait is, in theory, measured. The composite scores for the whole task are disproportionately distributed between the two scoring instruments.

As for the issue of assessing English proficiency based on a Hungarian text, the production-based criteria for correctness only measure English proficiency in an indirect manner and to a negligible degree. Although the reception-based criteria would in principle be

adequate for measuring reading comprehension, allowing dictionary use for the task interferes, making them unreliable for assessment. The issues highlighted above are exacerbated by the fact that the mediation subtest constitutes 30% of the maximum points achievable on, and 50% of the points required to pass the writing examination and receive certification on B1 level of proficiency in English. To conclude, the information the task provides is inconsequential for making meaningful, relevant or sufficient interpretations to base certification decisions on, calling into question the validity of the whole of the examination because the arguments in the Assessment Use Argument are not convincingly substantiated.

The analysis of the BGE task revealed that the descriptions of language ability were general and ambiguously phrased, as well as insufficiently clear and comprehensive for intra- or interrater consistency. Although this did not adhere perfectly to the principle of stochastic independence, the consequences of this fault are much smaller compared to those in the Profex examination. As for assessing English proficiency, in this case, the production-based criteria for correctness were found to have no bearing on measuring it. Nevertheless, the reception-based criteria do provide relevant and meaningful information for interpretations about English reading comprehension. The points for this task proportionate 1:9 in terms of the maximum score for the oral examination, making the already fewer and less severe (than for the Profex task) faults of the BGE exam have a considerably smaller negative effect on the certification decisions made. Nonetheless, the validity of the examination is still to be questioned, because some of the arguments in the Assessment Use Argument are not convincingly substantiated.

References

Bachman, L. F., & Palmer, A. (1996). *Language testing in practice: Designing and developing useful language tests*. Oxford University Press.

Bachman, L. F., & Palmer, A. (2010). *Language assessment in practice. Developing language assessments and justifying their use in the real world.* Oxford University Press.

Benke, E. (2002). Amit a számok közvetítenek. In M. Feketéné Silye (Ed.), *Porta Lingua. Szaknyelvoktatásunk az EU kapujában* (pp. 62–70). SZOKOE. *Debreceni Egyetem*. http://szokoe.hu/porta-lingua/archivum/porta-lingua-2002

BGE Language Examination Centre (2025). Results calculation for bilingual examinations. nyelvvizsgak.hu/page/default.asp?id=YWFJCH

Council of Europe. (2001). *Common European framework of reference for languages: Learning, teaching, assessment*. Council of Europe.

Council of Europe. (2020). *Common European framework of reference for languages: Learning, teaching, assessment. Companion volume*. Council of Europe.

Educational Authority (2025a). Accreditation centre for foreign language examinations. https://nyak.oh.gov.hu/default-eng.asp

Educational Authority (2025b). *Accreditation manual*. Accreditation Centre for Foreign Language Examinations. Akkreditacios_Kezikonyv_2024.pdf

Fulcher, G., & Davidson, F. (2007). *Language testing and assessment: An advanced resource book*. Routledge.

Jakobson, R. (1959). On linguistic aspects of translation. In L. Venuti (Ed.), *The translation studies reader* (pp. 113-118). Routledge.

Kaindl, K. (2013). Multimodality and translation. In C. Millán & F. Bartrina (Eds.), *The Routledge handbook of translation studies* (pp. 257–269). Routledge.

Tankó, Gy. (2005). *Into Europe: Prepare for modern English exams – The writing handbook*. Teleki László Alapítvány.

Tankó, G. (2019). A validálási folyamat érvelésalapú megközelítéseinek áttekintése. *Modern Nyelvoktatás*, *25*(3–4), 65–85.

Tankó, Gy. (2022) *Paraphrasing, summarising and synthesising skills for academic writers: Theory and practice* (3rd ed). Eötvös University Press.

Toulmin, S. E. (1958). *The uses of argument*. Cambridge University Press.

Toulmin, S. E. (2003). *The uses of argument* (Updated ed.). Cambridge University Press.

Warta, V., Hegedűs, A., Kránicz, R., & Hambuchné dr. Kőhalmi, A. (2025). *PTE ÁOK és ETK: PROFEX orvosi szaknyelvi nyelvvizsgarendszer vizsgaleírás (specifikáció)*. orv_vizsgaleiras_specifikacio_2025_01_22 vgl_Tok.pdf

Appendices

Appendix A: Profex B1-level English for Medical Purposes examination

B1 szint, közvetítőkészség

Kijelölt szöveghelyek

Mind az 5 kijelölt szöveghely 0, 1 vagy 2 pontot ér. Összesen 5 x 2 = 10 pont szerezhető maximálisan.

- 2 pont akkor kerül levonásra, ha magyarul teljesen értelmetlen a megfogalmazás, ha az információt félreértették vagy hiányosan közvetítették, ha a szövegkohézió vagy a koherencia sérült
- 1 pont akkor kerül levonásra, ha a mondatszerkesztés nehézkes, de a hiba a közölt információ lényegét nem érinti pl. egy szó, vagy rövidebb szerkezet félreértése, rossz szóválasztás kimaradó nem lényeges szó

Kijelölt szöveghelyek:

Szöveghely	Pontszám	Indoklás
1.	2	
2.	2	
3.	0	Az információt teljesen félreértette a vizsgázó, és ezért annak átvitele nem valósult meg.
4.	1	A hiba a közölt információ lényegét nem érinti, ám a magyartalan megfogalmazás miatt az információátvitel sérül.
5.	2	

Összbenyomás

5 pont	• megfelelő szókincs és nyelvtani szerkezeteket használata • minden lényeges információ közvetítése esetleges kisebb hiányosságok, vagy pontatlanságok mellett
4 pont	• **szöveg lényegi értelmének közvetítése kisebb pontatlanságokkal, néhány grammatikailag helytelen szerkezettel vagy pontatlan szóhasználattal** • **kisebb információs sikkasztások, amelyek azonban a szövegkohéziót nem sértik**
3 pont	• gyakori fordítás pontatlanságok • néhány súlyos és több enyhe grammatikai hiba • a szövegkohézió 1-2 helyen sérült
2 pont	• gyakran sérült szövegkohézió • gyakori félrefordítások és pontatlanságok
1 pont	• a szöveg csak kis mértékben tartalmaz értékelhető, érthető vagyis helyes információt
0 pont	• nem értékelhető teljesítmény

PROFEX SZAKNYELVI VIZSGA **B1**(alapfok) – angol nyelv Írásban teljesítenő **Közvetítés**		**vizsgázó sorszáma:** ☐☐☐☐ **MINTATESZT**

Fordítsa le az alábbi szöveget! Elérhető pontszám: 15 pont.
Figyelem! A vizsga akkor lehet sikeres, ha a vizsgázó részegységenként legalább 40%-ot teljesít.
Végső megoldásként csak a tintával írt változatot fogadjuk el.

Diabetes

Diabetes is a disease in which your blood glucose, or blood sugar levels are too high. Glucose comes from the foods you eat. Insulin is a hormone that helps the glucose get into your cells to give them energy. With type 1 diabetes, your body does not make insulin. With type 2 diabetes, the more common type, your body does not make or use insulin well. Without enough insulin, the glucose stays in your blood. Over time, having too much glucose in your blood can cause serious problems. It can damage your eyes, kidneys, and nerves. Diabetes can also cause heart disease, stroke and even the need to remove a limb. Pregnant women can also get diabetes, called gestational diabetes. Blood tests can show if you have diabetes. Exercise and weight control can help control your diabetes. You should also monitor your blood glucose level and take medicine if prescribed. Why is it important to prevent, diagnose and treat diabetes?

Untreated diabetes can lead to a number of serious problems, including:

- Eye damage that can cause blindness
- Kidney failure
- Heart attacks
- Nerve and blood vessel damage
- Problems with gums, including tooth loss

That's why treatment is important at any age. Keeping blood sugar levels very close to the ideal can minimize, delay and, in some cases, even prevent the problems that diabetes can cause.

(Source: https://www.nlm.nih.gov/medlineplus/diabetes.html)

Sources:

A_ERT_B1_K.pdf

Fordítsa le az alábbi szöveget

Appendix B: BGE C1-level tourism and catering examination

KÉTNYELVŰ FELSŐFOK
ÍROTT SZAKMAI SZÖVEG FORDÍTÁSA MAGYAR NYELVRE

KER skálák

	szakmai értelmezés	közvetítés
C1	Minden részletében meg tudja érteni az olyan hosszú és összetett szövegek széles körét, amelyek előfordulhatnak a társadalmi, tanulmányi és szakmai életben, azonosítja az apróbb részleteket, beleértve az attitűdöket, valamint a burkolt és kifejtett véleményeket.	(...) ki tudja választani a megfelelő nyelvi formát, hogy világosan ki tudja fejezni magát anélkül, hogy korlátoznia kellene, amit mondani akar. Jól használja széles körű (szakmai) szókincsét, az esetleges hiányokat körülírásokkal könnyedén áthidalja; ritkán kell keresgélnie a kifejezéseket vagy elkerülési stratégiákat alkalmaznia. (...) Alkalmanként kisebb tévesztések, de semmi jelentős szóhasználati hiba.

Az értékelési szempontok sávleírása

	szakmai értelmezés	közvetítés
C1 MEG-FELELT	5 – a szintnek kiválóan megfelelt	
	a vizsgázó az információt részletesen és pontosan, minden árnyalatában tökéletesen értelmezi	a vizsgázó nyelvileg igényes, összefüggő magyar szöveget produkál, pontos szakkifejezésekkel, esetenként kompenzációs stratégiákkal
	4 – a szintnek jól megfelelt	
	az információ teljességét helyesen értelmezi, árnyalati tévedések előfordulnak	a vizsgázó összefüggő magyar szöveget produkál, többnyire pontos szakkifejezésekkel, esetenként kihagyással
	3 – a szintnek még megfelelt	
	az információ lényegét helyesen értelmezi	a vizsgázó nyelvileg még elfogadható, összefüggő magyar szöveget produkál, a szakkifejezések helyett néhol nem megfelelő vagy általános kifejezéseket használ
NEM FELELT MEG	2 – a szintnek némileg alatta maradt	
	egy-két nagyobb félreértés, kihagyás előfordul	a vizsgázó töredékes, összefüggéstelen magyar szöveget produkál / a megfogalmazás lassú, nehézkes, akadozó
	1 – nem felelt meg	
	több súlyos félreértés adódik, kihagyások jellemzőek	a közvetítés során alig derül ki valami a szöveg témájáról
	0 – teljesítménye értékelhetetlen	
	nem értékelhető	nem értékelhető

Part 4: Translate the following text into Hungarian.

The role of customer satisfaction

Hospitality and tourism have evolved into truly global industries in which both consumers and producers are dispersed worldwide. Due to changes in lifestyle (including changes in work patterns, travel needs, eating habits, and the development of a cosmopolitan community), the services offered by hospitality businesses are now considered to be necessities, rather than luxuries. Consequently, during the past decade, there has been an exponential growth in hospitality businesses to meet the demands of the growing market. This has provided consumers with a great variety of choices while simultaneously augmenting competition in the marketplace. Moreover, it has become increasingly difficult for firms to assume that there exists an unlimited customer base prepared to maintain patronage. Hence, in the scheme of business, it has become apparent that the ultimate goal of any organization in a hyper-competitive market, is to maintain a loyal customer base.

Sources:

szóKÉTNYELVŰ FELSŐFOK.pdf

MINTA_AIF_szobeli.pdf